Neighbor selection for cold users in collaborative filtering with positive-only feedback

Alejandro Bellogín[1], Ignacio Fernández-Tobías[2], Iván Cantador[1], and Paolo Tomeo[3]

[1] Universidad Autónoma de Madrid, 28049 Madrid, Spain
{alejandro.bellogin,ivan.cantador}@uam.es
[2] NTENT, 08018 Barcelona, Spain ifernandez@ntent.com
[3] Politecnico di Bari, 70125 Bari, Italy paolo.tomeo@poliba.it

Abstract. Recommender systems heavily rely on the availability of historical user preference data, struggling to provide relevant suggestions for new users. The cold start user scenario is thus recognized as one of the most challenging problems in the recommender systems research area. Previous work has focused on exploiting additional information about users and items –e.g., user personality and item metadata– to mitigate the lack of user feedback. However, it is still unclear how to approach the worst scenario where no side information is available to a recommender system. Addressing this problem, in this paper we focus on new users of memory-based collaborative filtering methods with positive-only feedback, and conduct a comprehensive study of a number of neighbor selection strategies. Specifically, we present empirical results on several datasets analyzing the effects of choosing adequately the user similarity, the set of candidate neighbors, and the size of the user neighborhoods. In particular, we show that even few but reliable neighbors lead to better recommendations than large neighborhoods where cold start users belong to.

Keywords: recommender systems · collaborative filtering · cold start · neighbor selection.

1 Introduction

Recommender systems are designed to predict the most potentially relevant unknown items for a particular user considering her historical information, such as her interaction with the system in the form of clicks, likes, ratings and so on. This task is highly difficult with new users, since for them the available information is not enough to properly understand their preferences. In fact, this problem, known as the *user cold start*, has been recognized as one of the most challenging problems in the recommender systems research area, and there is not a unique solution that can be generically applied [5, 9].

Previous work has focused on acquiring and exploiting additional information about users and items to mitigate the lack of user feedback, such as side

information [1, 6], item metadata [11], cross-domain data [4], elicited user preferences [7], and user personality [5]. In this work, we face the worst scenario where no side information is available to the recommender system, which has to rely only on historical information about the users.

Within the context of cold start, in [9] Kluver and Konstan presented an analysis of the behavior of different recommendation algorithms, including a user-based nearest-neighbor (kNN) method with a neighborhood size of 30. Their empirical comparison showed that kNN produces poor quality recommendations for new users, resulting the worst method in that case. Therefore, the authors proposed to better investigate this issue as future work.

In this paper, we address the problem of low recommendation quality of user-based kNN for cold users when positive-only feedback (e.g., likes, purchases, views) is available. Hence, a novelty of our study arises in the combination of such issues: cold start and positive-only feedback. Several authors have already studied nearest-neighbor algorithms [8, 2], but none of them have dealt explicitly with cold start situations. Moreover, in general, they have used data consisted of numeric ratings. In many social media, however, this type of user feedback is not common, and is replaced by unary or binary ratings, such as the *likes* of Facebook, Twitter and Instagram, and the *thumbs up/down* of YouTube, or by implicit user feedback, such as product views and purchases in Amazon, and music play counts in Spotify and Last.fm. Motivated by this situation, we aim to provide some guidelines about how user-based kNN algorithms should be exploited in cold start scenarios with positive-only feedback.

More specifically, on datasets from Facebook and Last.fm, we present a comprehensive study showing that neighborhoods of cold start users may be composed of other cold start users who negatively impact on the recommendation quality. Considering this, we investigate a number of strategies to select candidate neighbors, based on how the performance changes when considering all the users, only cold start users, and only warm (i.e., not cold start) users. For each strategy, we also evaluate several user similarity metrics and neighborhood sizes.

Our empirical results indicate that a compromise should be met between large neighborhoods of cold users (to promote diversity) and small neighborhoods of warm users (to achieve high accuracy). Moreover, the results show that some user similarity metrics (like Jaccard coefficient and Cosine) are more sensitive to these experimental configurations than others, namely the Overlap and Log-Likelihood metrics. We provide an analysis on the rationale behind these effects, and propose solutions to deal with them.

2 Case study

The main intuition behind this work is that there are cases where the number of neighbors chosen as input parameter for a kNN approach should not be the only variable to consider in order to produce positive variations on item relevance prediction results. We argue that a proper selection of "good quality" neighbors

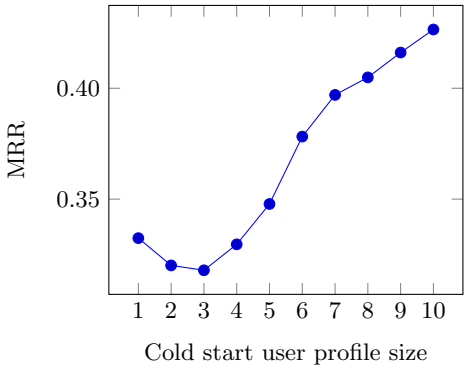

Fig. 1. User-based kNN with Jaccard and $k = 100$ (Facebook movie dataset).

may also positively affect the performance of a recommender system, especially when dealing with cold users having very few (even one) ratings in their profiles.

Let us consider the case represented in Figure 1, which shows the performance results (in terms of Mean Reciprocal Rank, MRR, that measures the inverse of the rank position of the first relevant item recommended) of a user-based kNN algorithm using the Jaccard coefficient as similarity metric, and 100 users as neighbors, on a dataset from Facebook with positive-only feedback. The performance of the algorithm decreases when we move from cold start user profiles containing 1 item to those with 2 and 3 items, whereas it increases with larger profiles. We observed this behavior in other datasets, such as those included in our experiments; the performance improves when one (music domain) or two (movie domain) additional interactions are introduced into the users' profiles after a performance drop that occurs when one interaction exists in the profiles.

We hypothesize that this effect on recommendation performance comes from a deficiency in how the used similarity metric selects the neighbors in cold start cases. For presentation purposes, we show the rationale behind our intuition by means of an analytic discussion about the Jaccard coefficient. The observations also hold for other similarity metrics used in kNN approaches.

Let u and v be two user profiles represented as vectors of positive-only feedback components in the item space, and let $|u|$ be the number of elements of u. The Jaccard coefficient (similarity) between u and v is defined as follows:

$$J(u, v) = \frac{|u \cap v|}{|u \cup v|} = \frac{|u \cap v|}{|u| + |v| - |u \cap v|} \tag{1}$$

As typical similarity metrics in the field, the Jaccard coefficient depends on the overlap $|u \cap v|$ between the two users involved.

It can be shown that for a target user, the larger a neighbor's profile, the more likely the probability of overlap X between the two users will be positive. This comes from the fact that the number of items common to the users can be

modeled as a hypergeometric distribution as:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{2}$$

being N the number of items, K the size of the target (cold start) user's profile, $n \leq N$ the size of the neighbor's profile , and $k \leq K$ the size of a target overlap.

For instance, if we take $k = K = 1$, then the probability that a random neighbor has an overlap of k is $\frac{n}{N}$, that is, the larger the neighbor's profile, the higher her chances of overlapping one item with the cold start user. In general, this applies when K grows, as there are more possibilities of having some (non-empty) overlap ($1 \leq k \leq K$), and those probabilities also depend on the neighbor's profile size.

One can argue that selecting larger neighbors is, in general, a good practice, since it provides more confidence on the resulting similarity, and thus allows for better recommendations; in fact, ad-hoc factors have been introduced into rating prediction to take this confidence into account when computing user similarity [8]. However, the Jaccard coefficient combines the overlap with the union of the users' profiles, hence providing a trade-off between the confidence on the similarity and the ratio of the overlap that belongs to each user; in other words, it does not promote larger neighbors anymore, but something in between, in order not to bias recommendations coming from heavy users. This can be shown by the following derivation:

$$J(u, v) = \frac{|u \cap v|}{|u| + |v| - |u \cap v|} = \frac{1}{\frac{|u|}{|u \cap v|} + \frac{|v|}{|u \cap v|} - 1} \tag{3}$$

In the context of cold start recommendation (where $|u \cap v| \leq |u|$ is very small), Equation 3 lets us conclude that those users v with large training profile sizes, and higher overlap probabilities (Equation 2), would generate very low values of the Jaccard similarity, whereas smaller users having some overlap with u would get higher similarity values. Actually, the optimal neighbor according to this metric is the one where $|u \cap v| = |u \cup v|$, since in that case $J(u, v) = 1$ (directly from Equation 1). However, these neighbors do not contribute with new items for the target user in recommendation time, and thus are useless in practice.

Because of this, and in order to assess our hypothesis that the observed drop in recommendation performance comes from the way the neighbors are selected by the similarity metric, in the next section, we present three strategies specifically tailored to the cold start scenario previously introduced. Furthermore, in the following sections, we analyze the impact of these strategies using different similarity metrics with small and large neighborhoods.

Unless stated otherwise, every experiment presented in this paper follows the methodology introduced in [9] to evaluate cold start scenarios, whose main characteristic is that every user being tested has the same number of items in training; in our experiments, from 1 to 10 items.

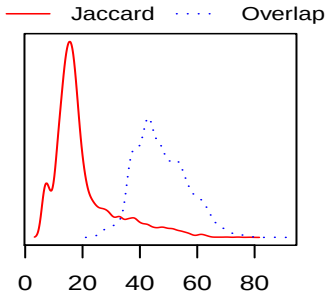

Fig. 2. Distributions of the average neighbor size using Jaccard coefficient and the overlap size (Facebook movie dataset). The former mostly selects neighbors with a very small profile, evidenced by the peak around 20 likes, whereas the latter shows a distribution with a moderate peak around 40 likes.

3 Strategies for neighbors selection

In the previous section, we hypothesized that a user-based kNN recommender using the Jaccard coefficient as similarity metric does not perform optimally due to the way it selects the neighbors. We noted that, although larger users have a higher probability of being selected because of their overlap, the Jaccard coefficient shows an inverse dependency on the neighbors' profile size. To further validate this hypothesis, in Figure 2 we show the distribution of the average neighbor size obtained when using the Jaccard coefficient as similarity metric, and compare it against the one obtained when pure overlap is used.

We observe that the average neighbor size when overlap is used as a similarity metric is much larger than when Jaccard coefficient is used. This is related to the previous observation that optimal neighbors for Jaccard are those small users who will not contribute with new items for test, and hence, no improvements can be expected from using such a similarity metric.

To mitigate this effect, we propose three strategies that impose constraints on the type of users that will be used in the neighborhoods:

All All users are equally considered to build the neighborhoods; this strategy is equivalent to the standard nearest-neighbor algorithm.

Cold Only cold start users are used to build the neighborhoods, that is, users with a profile larger than a threshold are not considered as potential neighbors.

Warm Cold start users are deleted from the neighborhoods, and only users with enough items (more than a given threshold) in their profiles are considered as neighbors.

The use of these neighbor selection strategies will allow us to analyze the effect that the type of neighbor (in terms of profile size) has in the final performance of the collaborative filtering method. Besides, by applying these strategies

with different similarity metrics, we study whether some similarities are more sensitive to these effects than others. Finally, we study the impact that the neighborhood size has on the recommendation performance for each strategy.

4 Experiments

In this section we detail the settings and results of the experiments conducted to evaluate the recommendation quality in cold start situations of the user-based kNN method on three datasets with positive-only feedback, namely two Facebook datasets used in [5] with page *likes* on the movie and music domains, and the HetRec 2011 dataset[4] with Last.fm music listening records.

4.1 Experimental settings

The conducted evaluation is based on a modified user-based 5-fold cross-validation methodology proposed in [9] for the cold start user scenario. First, we selected the users with at least 16 likes, shuffled and split into five (roughly) equally sized subsets. In each cross-validation stage, we kept all the likes from four of the groups in the training set, whereas the likes from the users in the fifth group were randomly split into three subsets to properly select the best configuration of the algorithms: training set (10 likes), validation set (5 likes), and testing (the remaining likes, at least 1). In order to simulate different user profile sizes from 1 to 10 likes, we repeated the training and the evaluation processes ten times, starting with the first like in the training set, and incrementally increasing it one by one feedback. This evaluation setting allowed us to evaluate each profile size with the same test set, avoiding potential biases in the evaluation, since some accuracy metrics have been proven to be sensitive to the test set size [9].

After preprocessing, the Facebook music dataset contained $49,369$ users, $5,748$ music bands and artists, and $2,084,462$ likes; the Facebook movie dataset contained $26,943$ users, $3,901$ movies, and $876,501$ likes. The Last.fm dataset contained $1,892$ users, $17,632$ artists, and $92,834$ relations between users and listened band/artist. Note that, although the Last.fm dataset includes listening counts, these are ignored and the information from the three datasets is considered as *unary feedback*, that is, only the signal that a user interacted with an item is used by the recommendation algorithms.

4.2 Results

In order to assess the ability of the strategies proposed in Section 3 to select good neighbors in collaborative filtering, and how this aspect affects the recommendation performance, we evaluated a user-based kNN method with a fixed value for the neighborhood size ($k = 100$), and varying its similarity metric and neighbor selection strategy.

[4] `https://grouplens.org/datasets/hetrec-2011`

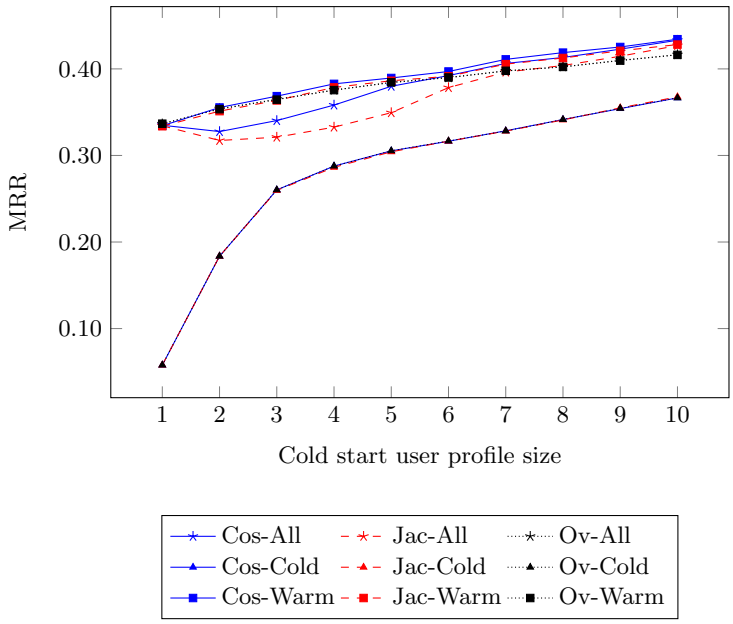

Fig. 3. MRR results in the Facebook movie dataset when $k = 100$.

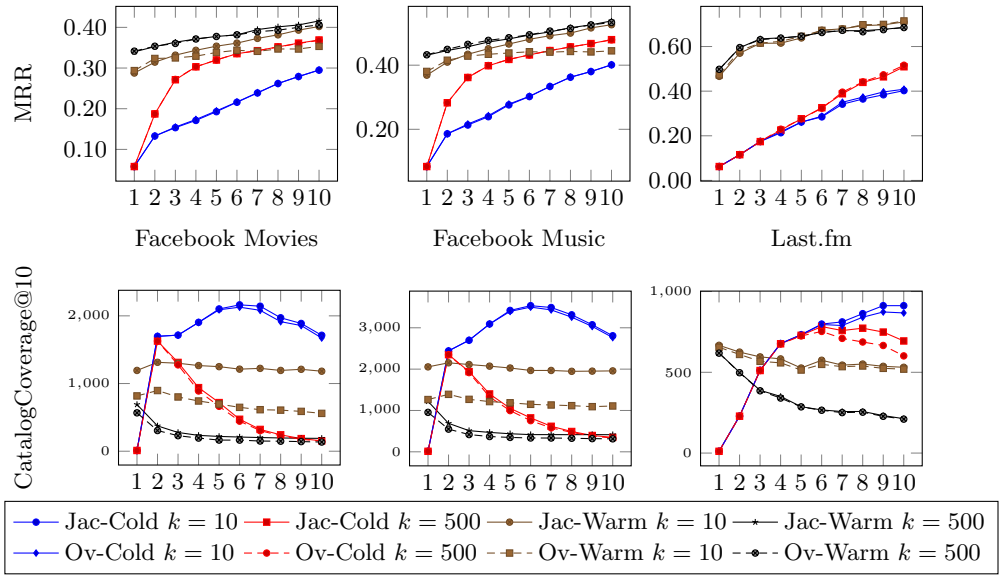

Fig. 4. Accuracy and diversity results at different neighborhood sizes (k). Cold start user profile size is represented in the X axis.

We show this comparison in Figure 3, where **Cos** denotes the Cosine similarity metric [10], **Jac** denotes the Jaccard coefficient presented in Section 2, and **Ov** is simply the user overlap, i.e., the amount of items in common between both users. The Log Likelihood ratio [3] was also tested, but produced the same results than Ov, and therefore it is not presented here. Note that since the used datasets contained positive-only feedback, if needed, we modified the similarities accordingly, e.g., by binarizing the input data.

In the figure we observe that the *Warm* strategy achieves always better performance values than other strategies, independently of the used similarity metric. This strategy exploits neighbors with larger profiles (in our case, those not considered as cold-users, i.e., with more than 16 items). Hence, we conclude that the size of the neighbors' profiles plays an important role in the recommendation stage, either because it allows producing item suggestions with higher confidence, or because its potential set of candidate recommended items is larger.

From Figure 3, we also note that the performance of the three similarities tested when the *Cold* strategy is applied is quite close, indicating that cold users are not beneficial as neighbors, no matter the similarity used. Nonetheless, the other two strategies make a difference; we observe that these similarities can be grouped depending on whether the *Warm* strategy obtains better or close results than *All*. For the Cosine and Jaccard metrics, the former case is true, since there is a clear difference between their performance, being larger for the *Warm* strategy. Overlap (and Log Likelihood ratio, actually), on the other hand, does not improve so much when using only warm users in the neighborhoods. This different behavior reinforces the observation raised in Section 2 with Figure 1: Overlap has an inherent bias towards incorporating users with larger profiles in their neighborhoods, and therefore it does not improve when cold users are filtered out. Additionally, Jaccard and Cosine similarity metrics tend to favor those neighbors with smaller profiles (as noted in Equation 3 for Jaccard), and, hence, when cold users are filtered out (*Warm* strategy) their performance improves.

To further understand the behavior of these similarities, and the effect of the proposed neighbor selection strategies, in Figure 4 we present results for two metrics, MRR for accuracy and CatalogCoverage@10 –measured as the percentage of items a method has recommended at least to one user– for diversity, on the datasets presented before. Here, we analyze two extreme neighborhood sizes: small ($k = 10$) and large ($k = 500$). The one presented in previous figures ($k = 100$) is in the middle of both. Also, considering the observations made about the facts that the *All* strategy is always outperformed by the *Warm* strategy, and that Cosine behaves in a similar way than Jaccard, we report results only for *Warm* and *Cold* strategies, using the Overlap and Jaccard similarity metrics.

The first issue to notice in this figure is that a tradeoff between accuracy and diversity is very clear in these experiments: worst recommenders in terms of MRR achieve the highest values of CatalogCoverage@10. In terms of neighborhood sizes, the tradeoff is also present, although less obvious in principle: small neighborhoods of large neighbors (*Warm* with $k = 10$) always outperform large neighborhoods of small neighbors (*Cold* with $k = 500$) in terms of accuracy.

However, depending on the dataset and the size of the cold user, the same effect occurs in terms of diversity.

For the Facebook movies and music datasets, large neighborhoods of small neighbors are more diverse than small neighborhoods of large neighbors when the target user is very cold (less than 3 observations in her profile). For the Last.fm dataset, in contrast, the situation is the other way around: recommendations are more diverse when more interactions are available in the cold start user profile. A possible explanation for this result could be attributed to the different nature of the datasets: whereas Facebook movies and music likes are explicit one-class user preferences (someone likes a page –and this is the only allowed interaction), Last.fm listening counts represent an implicit user feedback (the user's frequency shows the interest towards that item, but it is relative to the user's whole listening history). To some extent, this different nature may also affect how a cold user is defined in each system: someone with 10 page likes in Facebook is probably better represented than someone with 10 listening records in Last.fm. In any case, it is interesting to observe that the same gap in performance between warm and cold neighbors is found in both systems.

Finally, we also note that, independently of the neighbor selection strategy, the user similarity metric, and the size of the cold start user profile, the larger the neighborhood is, the better the accuracy the recommendation algorithm achieves. Hence, if computational resources (memory and processing time to generate the recommendations) are not a problem, the simplest solution to solve the accuracy problem in cold start scenarios would be to increase the size of the computed neighborhoods.

5 Conclusions and future work

In this paper, we have investigated the impact of some parameters in user-based nearest-neighbor algorithms when applied to cold start users with positive-only feedback. We have compared three strategies to build user neighborhoods – considering only cold users, excluding cold users from the neighborhood, and without any constraint–, and have observed that neighborhoods based on cold users are usually worse for accuracy, but may allow for more diversified recommendations, depending on the domain. Furthermore, our experiments showed that small neighborhoods of large neighbors outperform large neighborhoods of small neighbors, which might be used as a guideline for deploying kNN algorithms in cold start scenarios with positive-only feedback.

As future work, we plan to explore the behavior of the proposed neighbor selection strategies in a general setting where the ratio of cold users is unknown. Besides, according to our analysis on how the Jaccard coefficient finds optimal neighbors, we aim to study other formulations to improve its performance by, for example, filtering those neighbors with similarities equal or very close to 1.0, since they cannot contribute with new items for the target user.

Acknowledgments. This research was supported by the Spanish Ministry of Economy, Industry and Competitiveness (TIN2016-80630-P).

References

1. Barjasteh, I., Forsati, R., Masrour, F., Esfahanian, A.H., Radha, H.: Cold-start item and user recommendation with decoupled completion and transduction. In: Proceedings of the 9th ACM Conference on Recommender Systems. pp. 91–98. RecSys '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2792838.2800196
2. Bellogín, A., Castells, P., Cantador, I.: Neighbor selection and weighting in user-based collaborative filtering: A performance prediction approach. ACM Transactions on the Web **8**(2), 12:1–12:30 (Mar 2014). https://doi.org/10.1145/2579993
3. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics **19**(1), 61–74 (Mar 1993)
4. Enrich, M., Braunhofer, M., Ricci, F.: Cold-start management with cross-domain collaborative filtering and tags. In: E-Commerce and Web Technologies. pp. 101–112. Springer, Berlin, Heidelberg (2013)
5. Fernández-Tobías, I., Braunhofer, M., Elahi, M., Ricci, F., Cantador, I.: Alleviating the new user problem in collaborative filtering by exploiting personality information. User Modeling and User-Adapted Interaction **26**(2-3), 221–255 (2016). https://doi.org/10.1007/s11257-016-9172-z
6. Gantner, Z., Drumond, L., Freudenthaler, C., Rendle, S., Schmidt-Thieme, L.: Learning attribute-to-feature mappings for cold-start recommendations. In: Proceedings of the 2010 IEEE International Conference on Data Mining. pp. 176–185. ICDM '10, IEEE Computer Society, Washington, DC, USA (2010). https://doi.org/10.1109/ICDM.2010.129
7. Graus, M.P., Willemsen, M.C.: Improving the user experience during cold start through choice-based preference elicitation. In: Proceedings of the 9th ACM Conference on Recommender Systems. pp. 273–276. RecSys '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2792838.2799681
8. Herlocker, J., Konstan, J.A., Riedl, J.: An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. Information Retrieval **5**(4), 287–310 (Oct 2002). https://doi.org/10.1023/A:1020443909834
9. Kluver, D., Konstan, J.A.: Evaluating recommender behavior for new users. In: Proceedings of the 8th ACM Conference on Recommender Systems. pp. 121–128. RecSys '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2645710.2645742
10. Ning, X., Desrosiers, C., Karypis, G.: A Comprehensive Survey of Neighborhood-Based Recommendation Methods, pp. 37–76. Springer, Boston, MA (2015). https://doi.org/10.1007/978-1-4899-7637-6_2
11. Tomeo, P., Fernández-Tobías, I., Cantador, I., Noia, T.D.: Addressing the cold start with positive-only feedback through semantic-based recommendations. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **25**(Supplement-2), 57–78 (2017). https://doi.org/10.1142/S0218488517400116